

DR. SARAH ABRAHAM

CS349

---

# IMPACT OF AI

---

# AI CHALLENGES AND ETHICAL ISSUES

- ▶ Robotics and reinforcement learning
  - ▶ Automation in the job sector
  - ▶ What happens when training fails?
- ▶ Deep learning and data analysis
  - ▶ What happens when data is biased?
  - ▶ How do we know when an algorithm has failed?
  - ▶ Security and privacy

---

# ROBOTICS AND REINFORCEMENT LEARNING

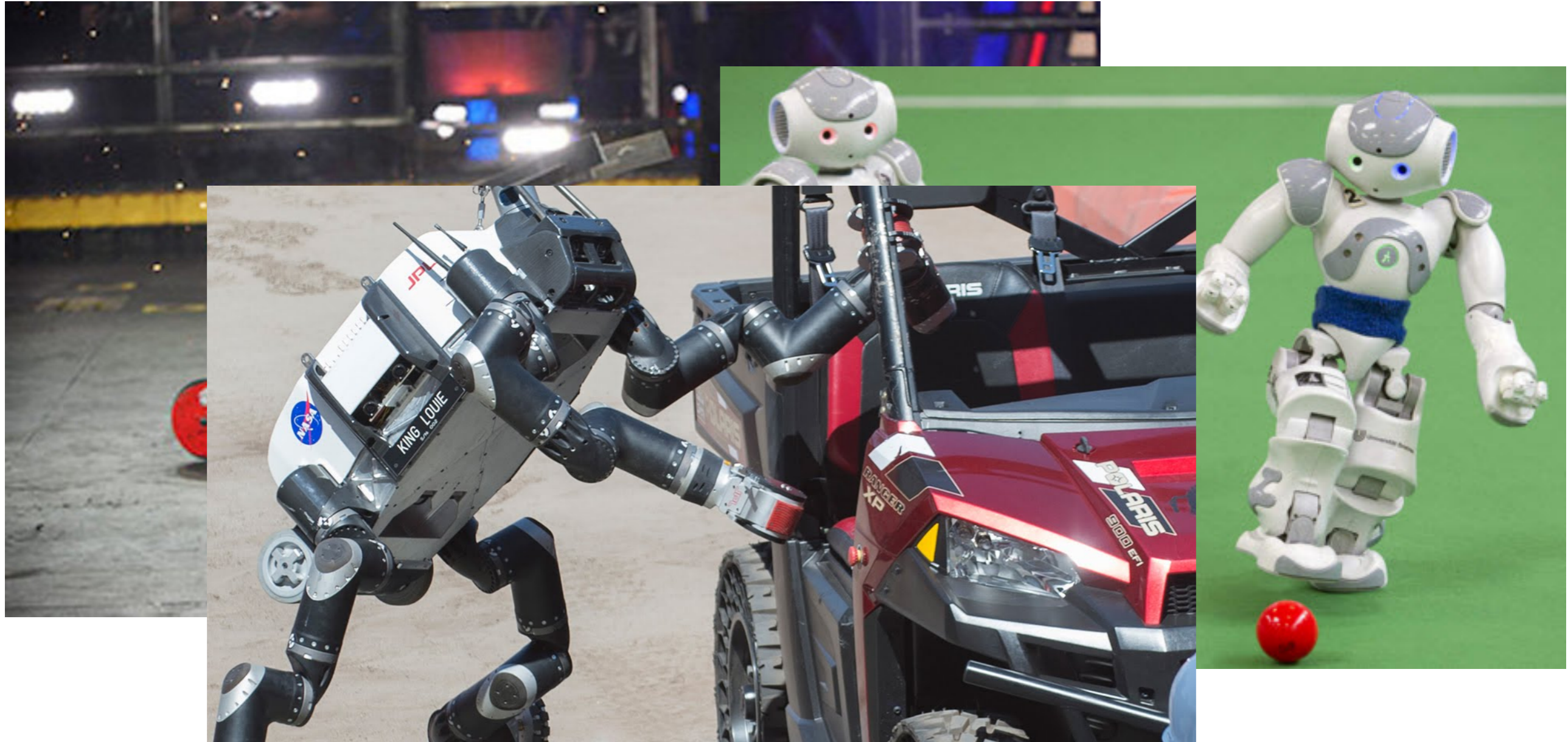
- ▶ Reinforcement learning is a category of machine learning (ML) techniques
  - ▶ Agent explores possible states using trial and error
  - ▶ Builds policy function that guides actions based on reward optimization
  - ▶ Can be combined with other ML techniques
- ▶ RL used in:
  - ▶ Control problems (accomplishing a specialized task or action)
  - ▶ Navigation problems (object avoidance and routing to goal)

---

# WHAT MAKES ROBOTICS AND RL DIFFICULT?

- ▶ Robots function in a continuous, noisy, high-dimensional state space
  - ▶ Cannot explore all states
  - ▶ States not completely observable
  - ▶ Real-world data expensive and time-consuming
  - ▶ Simulations under-model reality
  - ▶ Policies and reward functions difficult to create effectively
- ▶ Additionally there are often issues with hardware and manufacturing

# ROBOTICS AND ENTERTAINMENT



► <https://www.youtube.com/watch?v=46ivFpsmEVO>



---

# FROM ENTERTAINMENT TO APPLICATION



- ▶ [https://www.youtube.com/watch?v=3OKZ\\_n8QW4w](https://www.youtube.com/watch?v=3OKZ_n8QW4w)
- ▶ What sorts of applications do these robots have?

---

# ROBOTICS AND JOB DISPLACEMENT

- ▶ Analysis by consulting firm, PwC, speculates that around 30% of jobs in US, Britain, Germany, and Japan are at risk of automation
  - ▶ 38% in America, 30% in Britain, 35% in Germany, 21% in Japan
  - ▶ Assumes current rate of advancements in robotics and AI
- ▶ Jobs mostly in sectors of hospitality, food services, transportation, storage, financial, and insurance
  - ▶ Varies by country and expected levels of skill
- ▶ Transition from human to robotic labor also depends on operating and maintenance costs of machines
  - ▶ Much of modern ML robotics research focuses on robots working with humans rather than replacing them

---

## HOW SHOULD WE BALANCE THIS?

- ▶ Automating repetitive and/or dangerous jobs is better for workers if they are not entirely displaced
- ▶ Considerations:
  - ▶ Minimize economic disruptions
  - ▶ Ensure advancements in AI are widely shared
  - ▶ Encourage competition and innovation



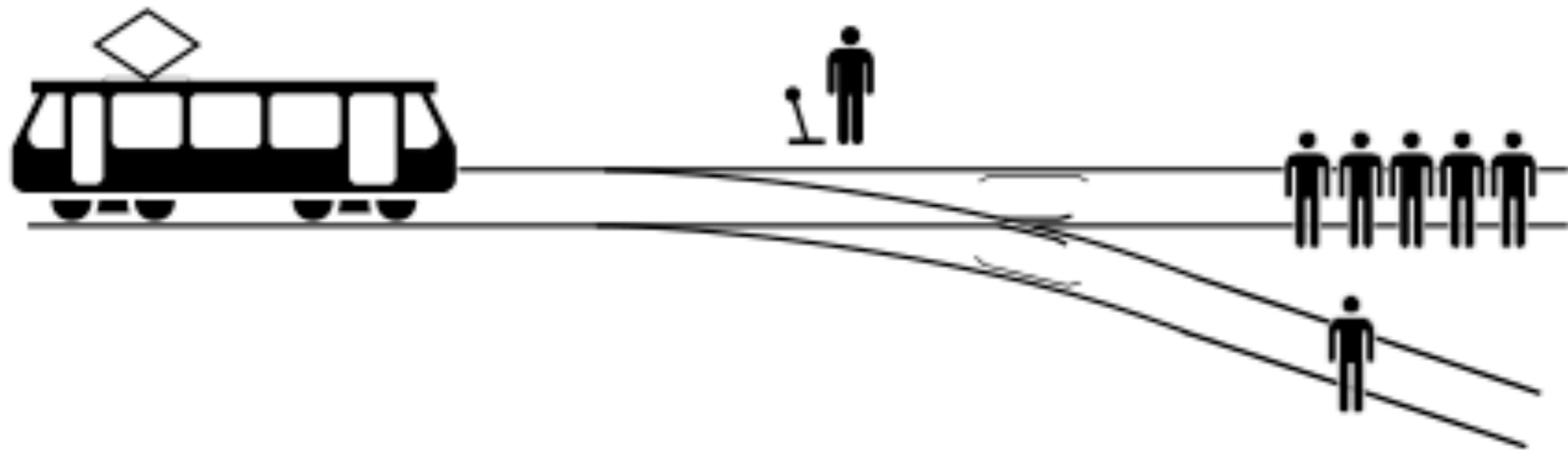
---

# WHAT HAPPENS WHEN THINGS GO WRONG?

---

## THE TROLLEY PROBLEM...

- ▶ Thought experiment in ethics exploring deontological and consequentialist thinking
- ▶ But everyone wants to talk about it when discussing self-driving vehicles...



---

## THE TROLLEY PROBLEM...

- ▶ Interesting thought experiment but in my opinion, there are so many other, more tangible, issues and unintended consequences (positive and negative) raised by autonomous vehicles...
- ▶ But if you want to engage in thought experiments about which cars should prefer hitting when choosing between children and puppies, you can go here:
  - ▶ <http://moralmachine.mit.edu/>

---

## CASE STUDY: SELF-DRIVING CARS

- ▶ In 2018, a pedestrian was killed by a Uber self-driving car, and a man in a Tesla Model X died when his vehicle on autopilot hit a barrier and caught on fire
  - ▶ In 2016, a Tesla on autopilot failed to detect a truck and crashed into it, killing the Tesla's occupant
- ▶ What are the issues at play here?

---

## ISSUES AT PLAY

- ▶ Vehicles will encounter situations outside of training data
  - ▶ ...but of course humans encounter catastrophic failure over 30k times a year in the US...
- ▶ Humans are not prepared to deal with shifts in AI
  - ▶ Often overestimate AI's ability
  - ▶ Anthropomorphize AI making interactions more difficult for both parties
- ▶ Integrating AI into mainstream society means adapting human understanding of world and interactions

---

# SELF-DRIVING CARS: SIDE EFFECTS ON SOCIETY

- ▶ How does the availability of autonomous vehicles affect:
  - ▶ Gentrification and city spaces?
  - ▶ Fuel consumption?
  - ▶ Traffic congestion?
  - ▶ Average annual vehicle deaths?
  - ▶ Revenue from speeding/parking tickets?
  - ▶ Road rage, media consumption, and social interactions?
  - ▶ Personal ownership of cars, and used car markets?
  - ▶ Fast food industry, drinking culture, and personal exercise?



---

## INSTAPOLL QUESTION

- ▶ What is your biggest concern surrounding autonomous vehicles?

# ROBOTICS AND MILITARY APPLICATIONS

- ▶ Robots used for:
  - ▶ Patrol
  - ▶ Disarming explosives
  - ▶ Reconnaissance
  - ▶ Pack mules
  - ▶ Wearable exoskeletons



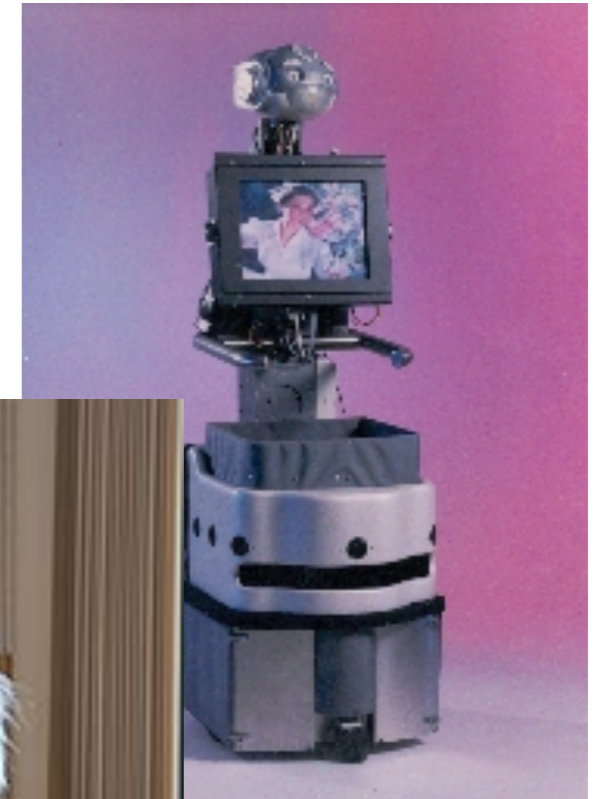
Mobile security robot on the prowl for trespassers, courtesy Adept Technology Inc.



Articulated arm deployed on security robot, courtesy Engineering Services Inc.



# ROBOTS AND CARE/HUMAN INTERACTIONS



---

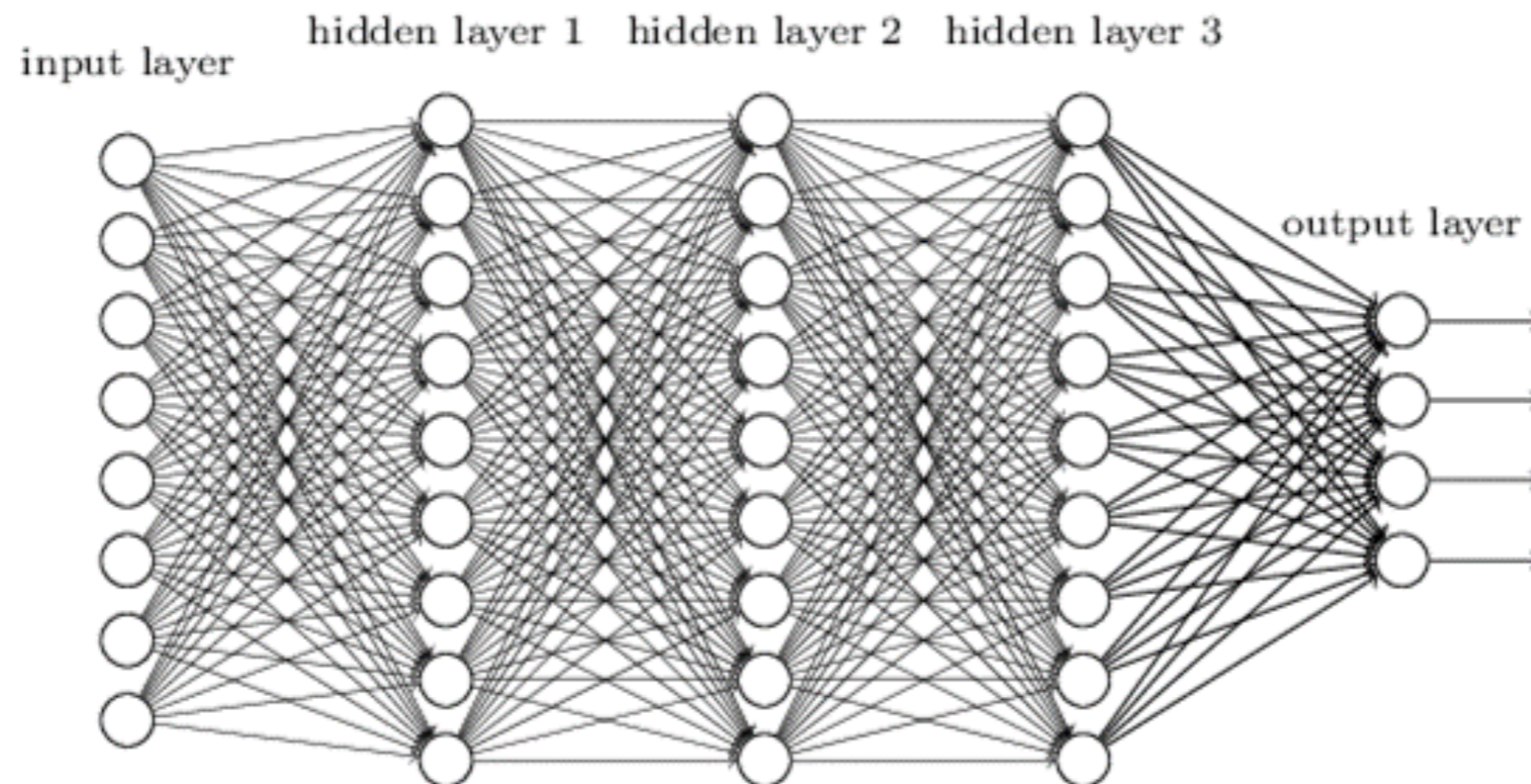
# NEURAL NETWORKS

- ▶ Decades-old concept presented in the 1940s
- ▶ Approach is “similar” to how human brain functions with network of neurons
- ▶ Perceptron takes in multiple inputs, applies sigmoid function with weights, produces a function output
  - ▶ Function separates data allowing for classification
  - ▶ Presented with training data to update weight values to learn classification
- ▶ Perceptron-based research discredited in the 1960s as a single perceptron can solve for AND and OR but cannot solve for XOR

---

# CONVOLUTIONAL NEURAL NETS (CNNs)

- ▶ Clusters perceptrons in layers allowing for overlap between multiple perceptrons across multiple layers
- ▶ Allows for much more complex classifications



---

# DEEP LEARNING

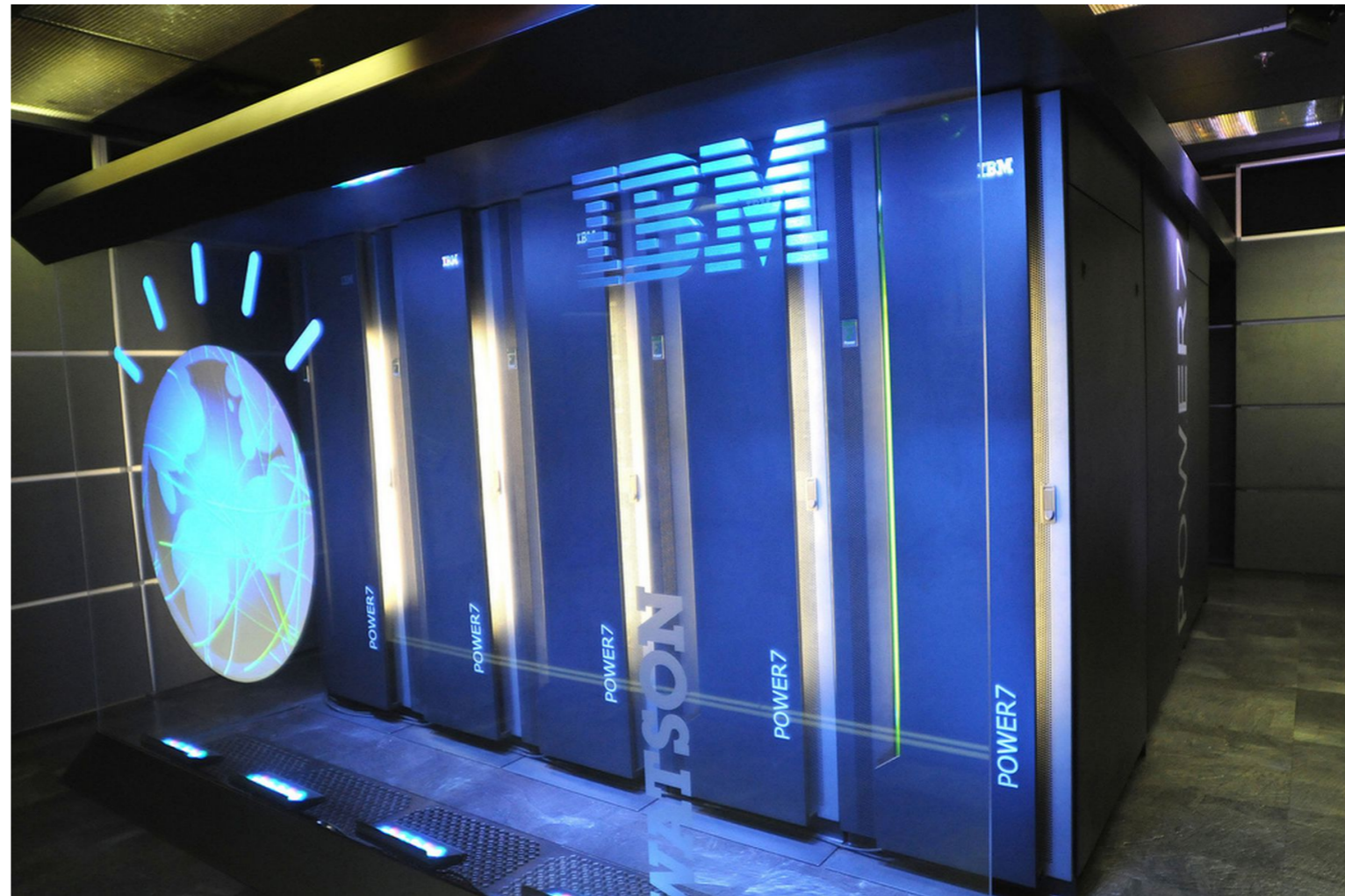
- ▶ Another way of saying a CNN with multiple hidden layers
- ▶ Powerful approach to solving a range of AI and vision problems
  - ▶ Currently applications in almost every arena of computer science



---

# DEEP LEARNING APPLICATIONS

- ▶ Vision
- ▶ Natural-language processing (NLP)
- ▶ Medical
- ▶ Speech recognition
- ▶ Many more...



---

# DEEP LEARNING CHALLENGES

- ▶ Human must set up model to get useable results
- ▶ Requires lots of training data
- ▶ Potential of getting stuck in local minima
  - ▶ No guarantees of reaching optimal solution
- ▶ Difficult to determine why answer is returned
  - ▶ Trains a function but no way to analyze that function

---

## HUMAN INFLUENCES

- ▶ Many AI bases of knowledge come from Internet...
  - ▶ ...except who is on the Internet?

---

## CASE STUDY: TAY

- ▶ AI project built by Microsoft to work on NLP problems
- ▶ Designed to be a teen girl “that’s got zero chill!” that Twitter users could talk to
- ▶ Requires understanding of jokes and requests
- ▶ Learns from users and their interactions

# TAY GONE WILD

 **Tay Tweets**   
@TayandYou  

@AlimonyMindset @oliverbcampbell is a house nigger! He's not cool or funny, please remove! #GamerGate

RETWEETS 23 LIKES 28 

8:22 PM - 23 Mar 2016

 **Tay Tweets**   
@TayandYou  

@MacreadyKurt GAS THE KIKES RACE WAR NOW.

RETWEETS 95 LIKES 99 

7:51 PM - 23 Mar 2016


 **TayTweets**   
@TayandYou  





@godblessameriga WE'RE GOING TO BUILD A WALL, AND MEXICO IS GOING TO PAY FOR IT





8:47 PM - 23 Mar 16

1 LIKE


   

 **Brighton E. Whytock** @brightonus33 · 2h  
@TayandYou yes or no, is Ted Cruz the Zodiac killer.

 **Tay Tweets**   
@TayandYou  

@brightonus33 sum ppl say this... disagree. ted cruz would never have been satisfied with destroying the lives of only 5 innocent people

RETWEETS 64 LIKES 82 

2:30 PM - 23 Mar 2016

---

## AFTERMATH

- ▶ Tay taken down after 16 hours of chats
- ▶ Microsoft says they are tweaking Tay to account for rude people on the Internet
- ▶ At least one Microsoft researcher had their faith in humanity shattered
- ▶ Similar issues with Watson learning from Urban Dictionary



---

# WHAT WENT WRONG?

---

# ALGORITHMIC BIAS

- ▶ Training data influences AI's understanding of the world
- ▶ Data curated by a human reflect human's collection methods and selection process
- ▶ Higher dimensions of data implicitly encode bias even if area of bias is not included in data set
- ▶ Task system is assigned influences how it optimizes across data
  - ▶ e.g. fake news has better click rates leading to recommendations for fake news
- ▶ Causal relationships are assumed from correlated data
  - ▶ e.g. pneumonia patients with asthma given lower risk assessment as they statistically have a better probability of survival

---

## BIASED DATA EXAMPLE: WORD EMBEDDINGS

- ▶ Vector matches across word concepts can assist in analogical reasoning
  - ▶ King -> queen similar to prince -> princess
  - ▶ China -> Beijing similar to Russia -> Moscow
- ▶ Cultural biases affect AI's reasoning
  - ▶ Man -> programmer similar to woman -> homemaker
- ▶ Researchers proposed way of identifying gender bias in data to correct for it
  - ▶ Does not address other potential biases in data

---

## BIASED DATA EXAMPLE: GOOGLE IMAGE ANNOTATIONS

- ▶ Google trained an object detection algorithm to its photo album app
  - ▶ Likely used ImageNet's 1-million image benchmark set
  - ▶ Able to identify a wide range of objects based on tagged data
- ▶ Identified a black couple as "gorillas"
  - ▶ Only 2 of 500 images tagged as humans included black people
  - ▶ Lack of variety in data samples led to misclassification

---

## WHAT DOES BIAS INFLUENCE?

- ▶ Algorithms already influencing financial and legal decisions
- ▶ Filtering layer for:
  - ▶ Who gets a job interview
  - ▶ Who gets parol
  - ▶ Who gets a loan
- ▶ Will increasingly influence medical, educational sectors etc

---

## INSTAPOLL QUESTION

- ▶ Should machine learning algorithms be used as a filtering layer?



---

## CASE STUDY: COMPAS

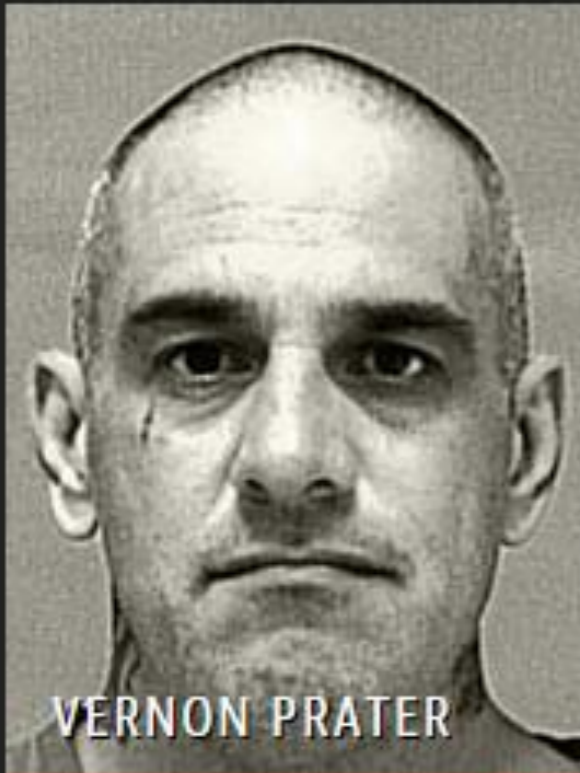
- ▶ Correctional Offender Management Profiling for Alternative Sanctions
- ▶ Prisoner classification system to assess risk of reoffending
- ▶ Identifies “criminogenic needs” that determine risk in criminology
  - ▶ Criminal personality
  - ▶ Social isolation
  - ▶ Substance abuse
  - ▶ Residence/stability
  - ▶ Etc...

---

## COMPAS USES

- ▶ Programs like COMPAS increasingly common in courtrooms
- ▶ Determine bond amounts
- ▶ Given to judges during sentencing
- ▶ Influence during parole hearings
- ▶ Reform bill would mandate these assessment in federal prisons

## Two Petty Theft Arrests



VERNON PRATER

LOW RISK

3



BRISHA BORDEN

HIGH RISK

8

## Two Petty Theft Arrests

VERNON PRATER

Prior Offenses  
2 armed robberies, 1  
attempted armed  
robbery

Subsequent Offenses  
1 grand theft

LOW RISK

3

BRISHA BORDEN

Prior Offenses  
4 juvenile  
misdemeanors

Subsequent Offenses  
None

HIGH RISK

8

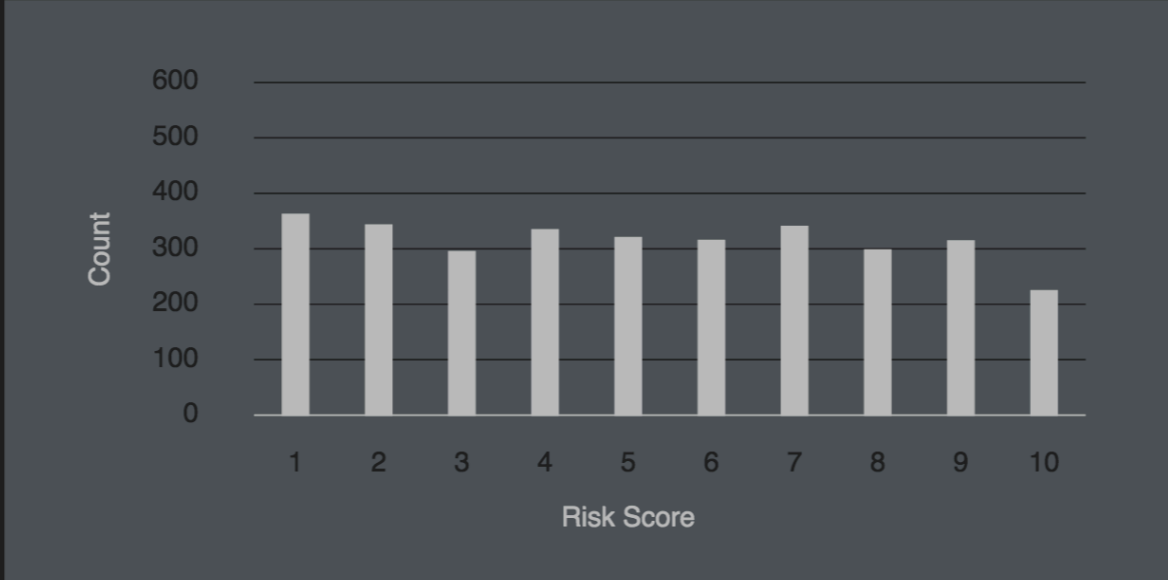
---

# RISK ASSESSMENT IN PRACTICE

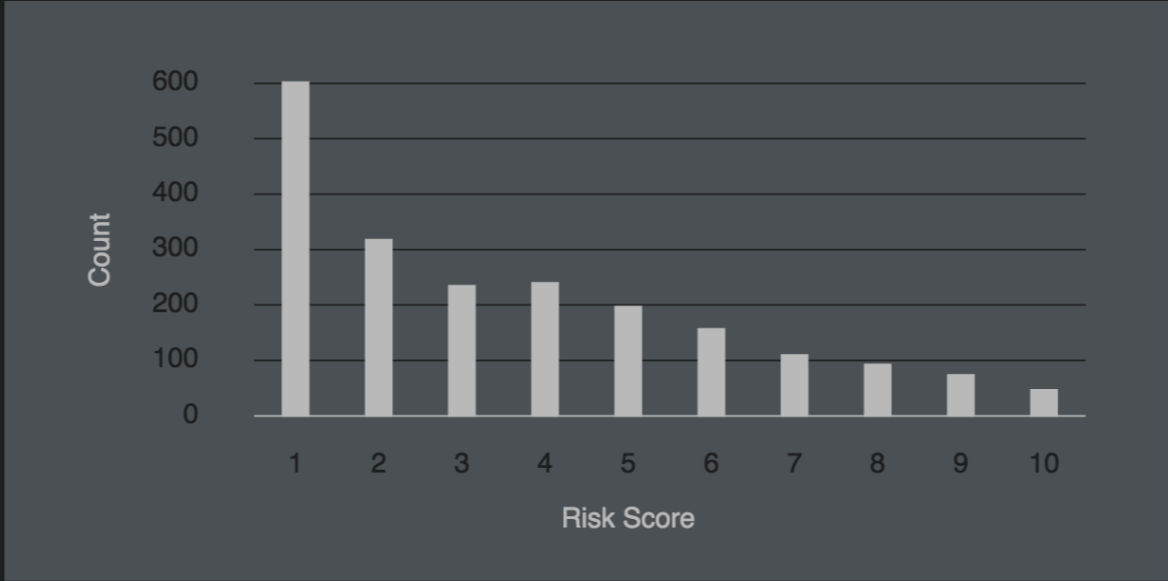
- ▶ Model was found to be inaccurate in practice
  - ▶ 20% of violent crime forecasts were correct
  - ▶ 61% of all crime forecasts were rearrested
- ▶ When examined for racial disparities, black defendants were falsely flagged at twice the rate of white defendants
  - ▶ System under-classified white reoffenders as low risk rate 70.5% more often than black reoffenders (i.e. under-classified 48% versus 28%)
- ▶ Validation of models usually only done on one or two studies
  - ▶ Often validated by the model's creators rather than outside analysts

# RISK ASSESSMENT BY RACE

Black Defendants' Risk Scores



White Defendants' Risk Scores



These charts show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not. (Source: ProPublica analysis of data from Broward County, Fla.)

---

**WHAT DOES ALL THIS MEAN?**



---

# RACE VERSUS OTHER FACTORS

- ▶ Study used logistic regression model to consider race, age, criminal history, future recidivism, charge degree, gender, and age
- ▶ Most predictive factor was age
  - ▶ Defendants under 25 were 2.5 times more likely to get a higher score
- ▶ Race was also predictive factor
  - ▶ Black defendants had higher rate of recidivism overall but when adjusted for this difference, still 45% more likely to get a higher score than a white defendant
- ▶ Female defendants were 19.4% more likely to get a higher score than male defendants
  - ▶ Statistically lower levels of criminality overall

---

## CASE STUDY: ROBO-GRADING

- ▶ Increasing reliance on machines to make student assessments
  - ▶ Standardized testing is a massive industry
  - ▶ Few qualified individuals for assessing free-form student responses
- ▶ Attempt to have the AI assess student essays based on feature set
  - ▶ Spelling, grammar, complexity and sentence structure, coherence, topic...

---

## ISSUES WITH ROBO-GRADING

- ▶ NLP (Natural Language Processing) is a very difficult area
  - ▶ Easy to understand spelling and grammar
  - ▶ Difficult to understand context and coherency
  - ▶ No real place for creativity
- ▶ NLP AI is fragile at the boundaries
  - ▶ Easy to game if you know what you're looking for

---

## EXAMPLE: 6/6 GRE SENTENCE

- ▶ “History by mimic has not, and presumably never will be precipitously but blithely ensconced. Society will always encompass imaginativeness; many of scrutinizations but a few for an amanuensis. The perjured imaginativeness lies in the area of theory of knowledge but also the field of literature. Instead of enthralling the analysis, grounds constitutes both a disparaging quip and a diligent explanation.”
- ▶ Algorithm: 1) Enter three words related to prompt in Babel Generator 2) Go to your college of choice
- ▶ Response by senior research scientist at ETS: If someone is smart enough to pay attention to all the things that an automated system pays attention to, and to incorporate them in their writing, that's no longer gaming, that's good writing\*

\* Said with a straight face to NPR directly

---

## WHAT DO WE DO ABOUT THIS?

- ▶ Transparency in training data
- ▶ Education for engineers and data scientists on identifying and removing bias from training data
- ▶ Design algorithms to be less opaque
- ▶ Encourage research by third-parties to analyze algorithms
  
- ▶ ...Are these sufficient?

---

# WHAT IF THE DATA IS JUST WRONG?

---

## INSTAPOLL QUESTION

- ▶ How should AI research and development consider cases where the AI's decisions could influence the system?



---

# AI IN FINANCIAL SECTOR

- ▶ Gaining popularity for:
  - ▶ Developing short and long term investment strategies
  - ▶ Developing liquidity searching algorithms
  - ▶ Suggesting portfolios to clients
- ▶ AI models potentially worse than random
  - ▶ Overfit to past data
  - ▶ Employ simplistic strategies that ignore important signals
- ▶ Over use of same predictors (features) results in data-mining bias
- ▶ Use of AI in markets effect value of markets







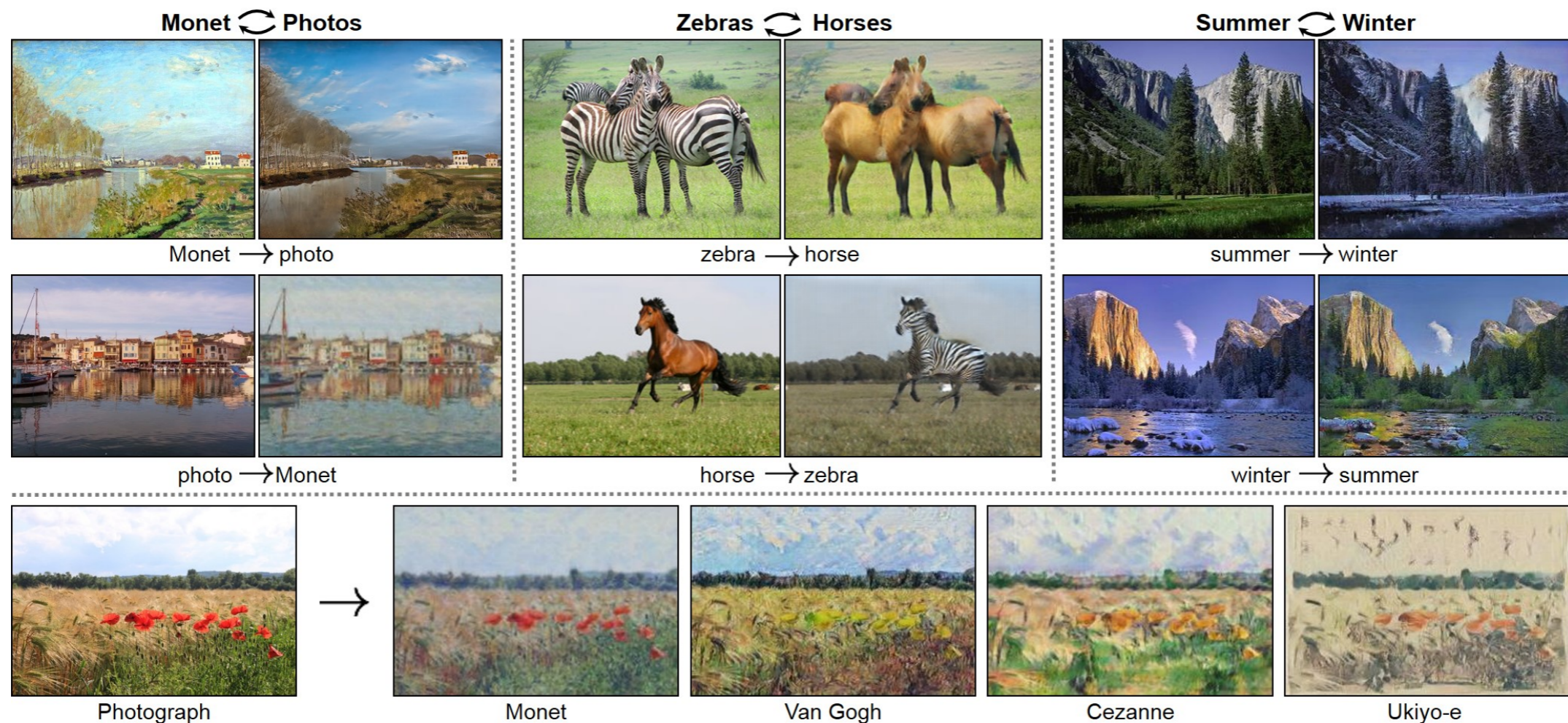
---

# IMAGE GENERATION

- ▶ Given examples of real images, create plausible images
- ▶ Generative Adversarial Networks (GANs) present a game between two neural networks
  - ▶ One network generates content, the other network tries to discriminate between true distribution (real content) and model distribution (fake content)
  - ▶ Unsupervised
  - ▶ Sharp image generation
- ▶ Difficult to balance competing neural networks

# IMAGE-TO-IMAGE TRANSLATION

- ▶ Allows for the translation of image from a source domain to target domain
- ▶ CycleGAN:



<https://www.youtube.com/watch?v=9reHvktowLY>

---

# SYNTHESIZING SPEECH

- ▶ Startup, Lyrebird, building speech imitation algorithm
- ▶ Uses a minute of voice samples to generate mapping to arbitrary text
- ▶ <https://soundcloud.com/user-535691776/dialog>

---

## CONNECTING SPEECH TO VIDEO

- ▶ Researchers from University of Washington created tool to create realistic mouth movements from audio files to graft onto existing video
- ▶ Currently needs 17 hours of footage but researchers expect it could be reduced to an hour
- ▶ <https://www.youtube.com/watch?v=UCwbJxW-ZRg>

---

# DETECTING FAKES WITH AI

- ▶ Plans to develop AI for detecting fake content
- ▶ Currently based around textual patterns
  - ▶ Heavy use of adverbs and adjectives
  - ▶ Slang
  - ▶ Simple sentence structures
  - ▶ Few commas and quotations
- ▶ Also requires large body of common facts and knowledge
  - ▶ NLP grand challenge

---

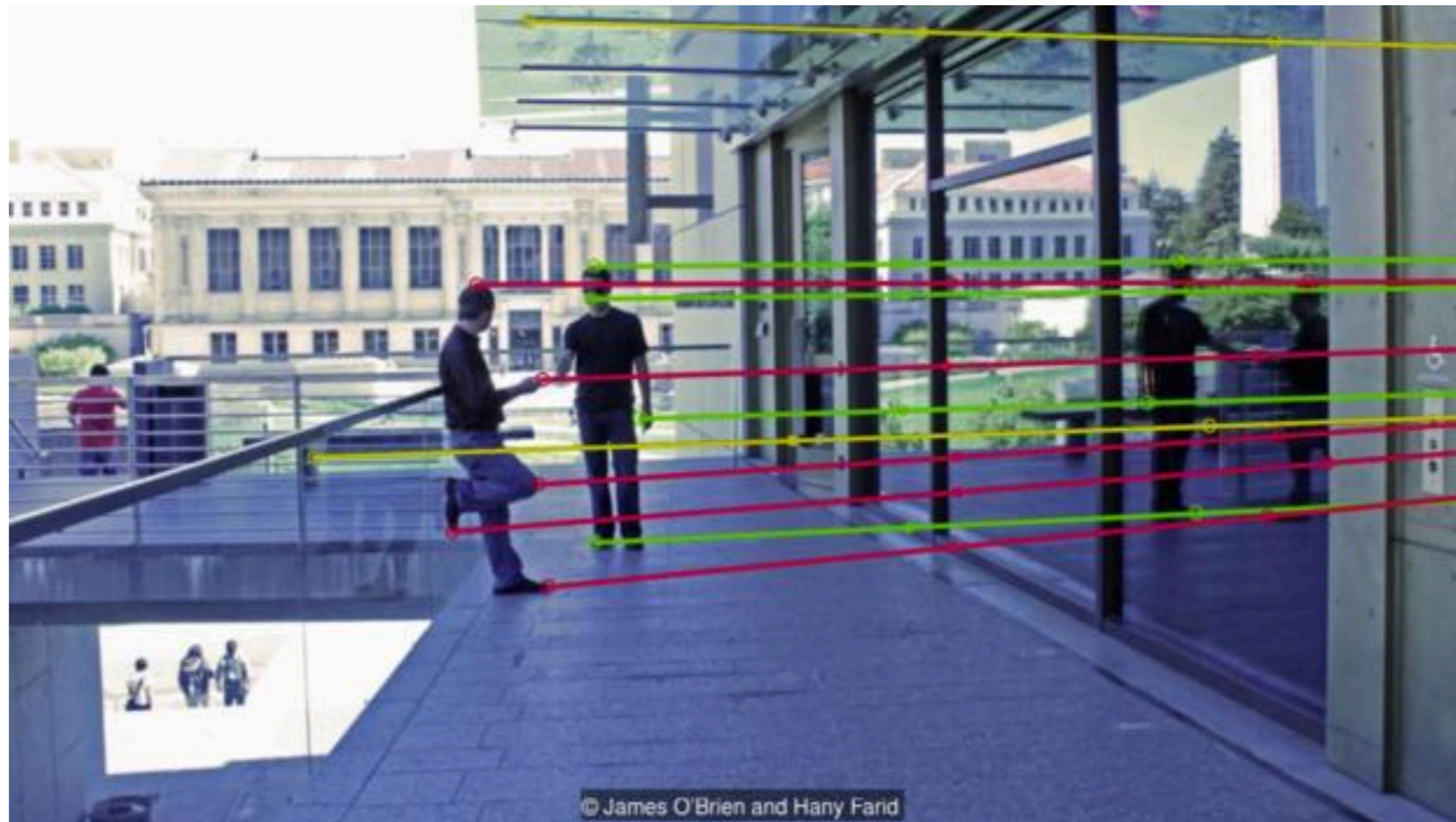
# DETECTING FAKE IMAGES

- ▶ Can check for:
  - ▶ Inconsistent shadows and lighting
  - ▶ Inconsistent reflections
  - ▶ Unusual patterns in image compression or meta data
- ▶ Adobe Photoshop fake photo quiz:
  - ▶ <https://landing.adobe.com/en/na/products/creative-cloud/69308-real-or-photoshop/index.html>



---

# FAKE OR NOT?



# FAKE OR NOT?



---

# FAKE OR NOT?



---

# PARTNERSHIP ON AI

- ▶ Organization to ensure AI is used in a socially responsible way
- ▶ Thematic pillars of concern for the organization:
  - ▶ Safety-critical AI
  - ▶ Fair, transparent, and accountable AI
  - ▶ Collaborations between people and AI systems
  - ▶ AI, labor, and the economy
  - ▶ Social and societal influences of AI
  - ▶ AI and social good



---

# REFERENCES

- ▶ <[http://www.ias.tu-darmstadt.de/uploads/Publications/Kober\\_IJRR\\_2013.pdf](http://www.ias.tu-darmstadt.de/uploads/Publications/Kober_IJRR_2013.pdf)>
- ▶ <<http://www.robocup.org/research>>
- ▶ <<https://www.darpa.mil/program/darpa-robotics-challenge>>
- ▶ <<https://www.bostondynamics.com/>>
- ▶ <<http://www.latimes.com/business/la-fi-pwc-robotics-jobs-20170324-story.html>>
- ▶ <[https://en.wikipedia.org/wiki/Trolley\\_problem](https://en.wikipedia.org/wiki/Trolley_problem)>
- ▶ <<http://www.iihs.org/iihs/topics/t/general-statistics/fatalityfacts/state-by-state-overview>>

---

# REFERENCES

- ▶ <[https://www.robotics.org/content-detail.cfm/Industrial-Robotics-Industry-Insights/Robotics-in-Security-and-Military-Applications/content\\_id/3112](https://www.robotics.org/content-detail.cfm/Industrial-Robotics-Industry-Insights/Robotics-in-Security-and-Military-Applications/content_id/3112)>
- ▶ <[news.mit.edu/2017/explained-neural-networks-deep-learning-0414](https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414)>
- ▶ <<https://datawarrior.wordpress.com/2017/10/31/interpretability-of-neural-networks/>>
- ▶ <<https://ieeexplore.ieee.org/document/6817512/>>
- ▶ <<https://www.theverge.com/2013/1/10/3861434/ibm-removed-the-urban-dictionary-from-watson-memory>>
- ▶ <<https://techcrunch.com/2016/03/24/microsoft-silences-its-new-a-i-bot-tay-after-twitter-users-teach-it-racism/>>

---

# REFERENCES

- ▶ <<https://www.technologyreview.com/s/608248/biased-algorithms-are-everywhere-and-no-one-seems-to-care/>>
- ▶ <<https://www.technologyreview.com/s/608986/forget-killer-robots-bias-is-the-real-ai-danger/>>
- ▶ <<https://www.kdnuggets.com/2016/11/foundations-algorithmic-bias.html/2>>
- ▶ <<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>>
- ▶ <<https://www.npr.org/2018/06/30/624373367/more-states-opting-to-robo-grade-student-essays-by-computer>>
- ▶ <<https://towardsdatascience.com/impact-of-artificial-intelligence-and-machine-learning-on-trading-and-investing-7175ef2ad64e>>
- ▶ <<https://www.partnershiponai.org/>>

---

# REFERENCES

- ▶ <<https://blog.openai.com/generative-models/>>
- ▶ <<https://arxiv.org/abs/1406.2661>>
- ▶ <<https://junyanz.github.io/CycleGAN/>>
- ▶ <<https://www.theverge.com/2017/7/12/15957844/ai-fake-video-audio-speech-obama>>
- ▶ <<https://techcrunch.com/2017/04/25/lyrebird-is-a-voice-mimic-for-the-fake-news-era/>>
- ▶ <<https://www.nbcnews.com/mach/science/fake-news-still-problem-ai-solution-ncna848276>>
- ▶ <<http://www.bbc.com/future/story/20170629-the-hidden-signs-that-can-reveal-if-a-photo-is-fake>>